Version 6.0 // Jan 2020

Morals From Rationality Alone? Some Doubts

J.P. Messina and David Wiens

Abstract. Contractarians aim to derive moral principles from the dictates of instrumental rationality alone. But it is well-known that contractarian moral theories struggle to identify normative principles that are both uniquely rational and morally compelling. Michael Moehler's recent book, *Minimal Morality*, seeks to avoid these difficulties by developing a novel "two-level" social contract theory, which restricts the scope of contractarian morality to cases of deep and persistent moral disagreement. Yet Moehler remains ambitious, arguing that a restricted version of Kant's categorical imperative is a uniquely rational principle of conflict resolution. We develop a formal model of Moehler's informal game-theoretic argument, which reconstructs a valid argument for Moehler's conclusion. This model, in turn, enables us to expose how a successful argument for Moehler's contractarian principle rests on assumptions that can only be justified by subtle yet significant departures from the standard conception of rationality. We thus extend our understanding of familiar contractarian difficulties by showing how they arise even if we restrict the scope of contractarian morality to a domain where its application seems both promising and necessary. We show that the problem lies not in contractarians' immodest ambitions but in the modest resources rationality can offer to satisfy them.

Social contract theory aims to reconcile the demands of social morality with the dictates of rational prudence. The most austere strand of this tradition — exemplified by Thomas Hobbes (1985), James Buchanan (2000), and David Gauthier (1986) — tries to derive moral principles from the dictates of instrumental rationality alone, eschewing appeals to controversial moral norms such as autonomy, fairness, or equality. For these contractarian theorists, a principle P is justified if and only if rational individuals, consulting solely their private interests as they see them, would agree to restrain their conduct in accord with P rather than allow their conflicts to persist unchecked.

The latest installment in this tradition — Michael Moehler's *Minimal Morality* — is both innovative and ambitious. The innovation is a eqtwo-level theory of the social contract, which restricts the scope of contractarian morality to cases of deep and persistent

Authors' note. J.P. Messina is Assistant Professor of Philosophy (research) at the University of New Orleans; David Wiens is Associate Professor of Political Science at UC San Diego. We are grateful to Michael Moehler for extensive feedback on earlier versions of this paper. We also thank anonymous referees and an Associate Editor for *PPE* for helpful comments.

To appear in *Politics, Philosophy, and Economics*. Penultimate version (uncorrected).

moral disagreement, where its application is both promising and necessary as a basis for peaceful conflict resolution (2018, 14–22).[1] Using his two-level theory, Moehler aims to show that rational and self-interested individuals who radically disagree about moral matters can nonetheless agree to resolve their conflicts in accord with a unique normative principle — the "Weak Principle of Universalization" — that expresses thick Kantian moral ideals such as autonomy, equality, and impartiality (see, e.g., 6, 20, 95, 107, 126–139). The ambition is to come "as close to a Kantian conclusion as is possible within a Hobbesian moral framework" (139).

In this essay, we show that Moehler's argument for the Weak Principle of Universalization (WPU) is not as morally minimal as he claims, even given his significant restrictions on its scope of application. We start by presenting a formal model that reconstructs Moehler's informal game-theoretic argument for the claim that agreement on the WPU is uniquely rational in the relevant cases of conflict. This model presents a valid argument for Moehler's claim. It also enables us to expose how a successful argument for the WPU calls on normative resources beyond those derived from the standard (Bayesian) conception of instrumental rationality. The upshot of our argument through section 3 is that Moehler faces a choice: he can prescribe the WPU as a principle of conflict resolution but grant that it cannot be justified on rational grounds alone, or he can persist in looking for a unique principle that can be justified solely by appeal to rational prudence but concede that the WPU cannot fill this role. In either case, Moehler's innovative "two-level theory" is inadequate to meet his contractarian ambition. In section 4, we consider a revised argument on Moehler's behalf, according to which the WPU is rationally justified as an alternative to the Nash bargaining solution. Although this claim is less ambitious than Moehler's claim that coordinating on the WPU is uniquely rational, it should nonetheless hold considerable interest for contractarians, since the Nash bargaining solution often serves as a starting point for contractarian normative analysis. We show, however, that this argument fails too.

We conclude that Moehler fails to derive an attractive normative principle on strictly rational grounds. Perhaps this conclusion will be unsurprising. It is well-known that contractarian theories generally struggle to identify principles of political morality that are both normatively appealing and uniquely consistent with the dictates of rational prudence (e.g., Gaus 2011, chap. 2; Sugden 1990). Moehler restricts the scope of contractarian morality to cases of deep and persistent conflict precisely to overcome these long-standing challenges (20–2; also Moehler 2014, forthcoming). We thus extend our understanding of these familiar challenges by showing that they arise even if we accept Moehler's apparently

1 Bare page numbers in the text refer to Moehler (2018).

promising scope restrictions. We show, in other words, that the problem lies not in contractarians' immodest ambitions but in the modest resources instrumental rationality has to offer normative theorists.

1. MOTIVATING *MINIMAL MORALITY*

For social contract theorists, the central problem is to identify normative principles that can provide a basis for stable and peaceful cooperation among people with diverse and often conflicting ends. Moehler follows other Hobbesian contractarians in trying to justify such principles solely by appeal to individuals' instrumental rationality. Not that he thinks normative principles can never be justified otherwise. Where people share a commitment to thick moral ideals (such as a conception of fairness, reciprocity, or social cohesion), they have sufficient reason to comply with robust moral principles. But arguments from controversial moral ideals are not appropriate for circumstances where "agents may hold irreconcilable moral ideals, or irreconcilable interpretations of such ideals" (1). Where deep and persistent moral disagreements block the emergence of shared moral ideals, we are left to justify normative principles by appeal to people's shared commitment to rational prudence (see 14–22). So Moehler's contractarian argument is embedded within a "two-level contract theory", applying only to what Moehler calls "cases of conflict in the strict sense defined" (14). Such situations ("cases of conflict" hereafter) are characterized by two features. First, no social conventions exist for resolving conflict among the parties involved, and their moral disagreements prevent them from resolving their conflict by appeal to the moral principles they personally endorse or by appeal to impartial arbitration. Thus normative principles must be justified solely by appeal to the rational prudence of those involved. Second, the conflict is "high-stakes" in the sense that, in the absence of a peaceful resolution, the conflicting parties are prepared to use violence to pursue their disparate interests (15). Hence coordinating on a second-level principle is urgent.

Moehler argues that people involved in conflict so defined ought to resolve their disagreement by means of the WPU, which reads thus:

> In cases of conflict, only pursue your interests subject to the constraints that your opponents can (i) enter the process of conflict resolution at least from their minimum standards of living, if the goods that are in dispute permit it, and (ii) fulfill their interests above this level according to their relative bargaining power. (125)

To get an intuition for what the WPU requires, we note that it is meant to be a "stabilized" version of the classic Nash bargaining solution (NBS) (85–87, 92). The NBS requires that conflicts be resolved as a function of what parties to the conflict expect to get should they fail to find a cooperative solution, so that a person's share of the benefits of cooperation is proportional to their prospects in the case of persistent conflict. Moehler follows the convention of interpreting these "disagreement payoffs" as indicating the parties' relative bargaining power.[2] Accordingly, the NBS is a principle of conflict resolution that "assigns each agent a share of the gains of cooperation that is proportional to their actual relative bargaining power in distributive conflicts" (85).[3] Although the NBS "represents a natural agreement point for rational agents" in cases of conflict (56; see also 2010, 447, 453), Moehler thinks that rational people would ultimately find coordination on the NBS unsatisfactory because it does not guarantee that their basic needs will be satisfied. According to Moehler, agents who are left with an inadequate share of resources have an incentive to use violence to take what they can, which undermines peaceful cooperation (115–16). The WPU is supposed to stabilize cooperation by requiring that each person whose disagreement payoff would leave them unable to meet their basic needs be credited enough to assure basic needs provision prior to calculating each person's share of the benefits of cooperation, with the remainder being divided according to the NBS assuming these adjusted disagreement payoffs (cf. 115). (We explain the relationship between the NBS and the WPU more precisely in section 4.)

According to Moehler, the WPU is the *uniquely* justified principle for resolving conflicts between purely instrumentally rational agents. Moehler's argument for this claim assumes a model of instrumental rationality he calls "*homo prudens*" (17, 96–107, original emphasis). According to this model, rational agents prefer peaceful cooperation on (nearly) any terms over violent non-cooperation (we discuss the "nearly" caveat below) (8, 101). Importantly, rational agents who insist on resolving conflict in accord with their most preferred moral principle, even at the cost of violent conflict, do not fit the *homo prudens* model and thus "fall outside the scope of the two-level contractarian theory" (18, 101–2).[4] Additionally, *homo prudens* satisfies the axioms of standard (Bayesian) rational choice theory such that his preferences over outcomes can be represented by an expected utility function (100-01).

Although he argues for the WPU from rational prudence alone, Moehler emphasizes

2 Binmore (1998, 78–82), however, warns against this interpretation.

3 We state the NBS more precisely in section 4.

4 We might wonder, then, whether Moehler's theory applies to cases of unreasonable disagreement. We do not pursue this issue here (but see Gaus 2019), noting it only as an invitation to clarify the types of conflict to which Moehler's theory is meant to apply.

that the principle's content expresses robust Kantian ideals of autonomy, equality, and impartiality (13, 95, 123). Rational agents commit to autonomy because coercively imposing terms of conflict resolution leads to an unstable peace, which subverts the end of achieving stable and peaceful cooperation (27); they commit to equality because "natural equality" ensures that even the weakest parties in a conflict are able to kill the strongest (13, 97). The principle displays its anti-egoist credentials insofar as it constrains the self-interested behavior of rational agents in cases of conflict much in the way that Kant's categorical imperative constrains the selection of maxims to those that are compatible with the moral law. So understood, Moehler's is an ambitious attempt to show that, when individuals can agree on little else, rational prudence commits them to a unique and normatively attractive principle for conflict resolution, provided only that they rationally prefer cooperation to violence.

2. A GAME-THEORETIC ARGUMENT FOR THE WPU

The WPU demands that each agent receive a share of material resources "according to her basic needs and above this level according to her relative bargaining power" (18). In support of this distributive principle, Moehler presents an informal game-theoretic argument based on an "empathetic contractor" thought experiment involving rational agents — agents who fit the *homo prudens* model — playing what he calls a "peace game". His primary task is to show that resolving conflict in accord with the WPU is the "the only rational strategy for the agents in the described peace game" (125, also 129). Whereas Moehler uses "game-theoretic concepts merely as a heuristic device to express [his] arguments more clearly" (108), we develop a formal reconstruction of his argument. Our more formal approach enables us to clarify the set up of a complex thought experiment with many moving parts, and disciplines our inferences as we reason from the initial set up through to the conclusion (cf. Rubinstein 2012, 16–20; Rodrik 2015, 29–36, 46). Indeed, our model enables us to show how Moehler's thought experiment crucially turns on several assumptions that take his reasoning beyond the standard requirements of instrumental rationality.[5]

In keeping with his informal use of game theory, Moehler does not explicitly specify the "peace game" he uses to represent rational agents' interactions. This leaves us to make several modeling choices on Moehler's behalf, which requires some interpretive work on

5 To avoid distractions, we wish to forestall a worry about our strict use of game theory in what follows. The worry is that, by analyzing Moehler's argument within the strictures of classical game theory, we are (uncharitably) holding him to standards that he rejects. We take up this concern in the concluding section.

our part. Our objective throughout is to model what we take to be the most plausible reconstruction of Moehler's thought experiment that is able to produce his conclusion as nearly as possible. To help frame our reconstruction, we note that we interpret the "peace game" as a repeated coordination game. Some might instead interpret Moehler as presenting a bargaining-theoretic argument for the WPU. This seems natural enough: The WPU is presented as an amendment to the classic Nash bargaining solution, and at times, Moehler's exposition suggests that rational agents are involved in some bargaining when settling on a principle of conflict resolution.[6] Nevertheless, we do not find strong support for a bargaining-theoretic interpretation, for at least two reasons. First, Moehler's most explicit statements on the "peace game" describe it as a repeated coordination game (108, 121, 129). Second, had Moehler intended to sketch a bargaining game, we would expect him to provide some guidance regarding the rules that structure the bargaining process — rules determining who gets to make offers when, when counteroffers can be made, and how the bargaining process concludes. Yet we find no guidance in Moehler's text regarding any bargaining protocol. Moreover, his remarks at one place suggest that there is no bargaining protocol for the peace game.[7] We thus model the "peace game" as a repeated coordination game and leave it to others to develop a bargaining-theoretic model of Moehler's argument. We highlight this interpretive choice at the outset to explain our emphasis on the theory of repeated games rather than bargaining theory.

With these preliminary remarks in place, we begin our reconstruction of Moehler's thought experiment. Instrumentally rational agents conform to the *homo prudens* model outlined above. Conflict among rational agents is modeled as a "peace game", in which two agents ("players") are engaged in a series of discrete interpersonal conflicts (106–7). Iterated conflict of this sort is often represented using a repeated Prisoner's Dilemma model. Moehler, however, rejects the Prisoner's Dilemma as an appropriate model for his purposes, opting instead to represent cases of conflict using a "simultaneous-move pure coordination game" (121; also 108–9). The peace game is thus a repeated coordination game: a single instance of conflict is modeled as a static coordination game (the "stage game") in which players must divide the benefits of cooperation for a particular interac-

6 E.g., "negotiations among the agents would break down" (121); "the bargaining process will break down among rational agents" (126).

7 "I employ noncooperative game theory for the derivation of the principle of conflict resolution, because in cases of conflict in the strict sense defined all attempts to resolve disputes through dialogue among the parties to a conflict and by appeal to a third party are assumed to have failed, and the parties now must find independently tacit agreement on a principle of conflict resolution *without further negotiations* as to how to play the 'peace game' and without being able to make binding agreements before the game is played" (108, emphasis added).

	ROW	COL	WPU
ROW	$V, 0$	d_R, d_C	d_R, d_C
COL	d_R, d_C	$0, V$	d_R, d_C
WPU	d_R, d_C	d_R, d_C	$W_R(V), W_C(V)$

Figure 1. Moehler's stage game

tion. Ongoing conflict is modeled as the repetition of the stage game over an unspecified number of time periods. A *principle of conflict resolution* instructs players how to act in each particular stage game. An *instrumentally rational* principle of conflict resolution is a principle that specifies a mutually rational strategy profile for the repeated coordination game. Roughly, a principle is *mutually rational* so long as no player has an incentive to unilaterally deviate from the dictates of the principle in any stage game. (We give a more precise definition of the relevant notion of mutual rationality below.) The players' task is to identify a (hopefully unique) rational principle of conflict resolution by considering their respective payoffs from various strategy profiles for the repeated game.

Figure 1 summarizes our reconstruction of Moehler's stage game. There are two players: one is associated with the rows of the matrix, the other with the columns. For convenience, we call these players Rowan and Collins.[8] The row and column headings specify the players' available strategies. A careful reading of Moehler's text immediately suggests three candidates. One is of course, the WPU; this is compared with two others, which we might call *give Rowan everything* (ROW) and, by the symmetry of the game, *give Collins everything* (COL) (122). Moehler's analysis leaves the impression that this list exhausts the principles available to agents in cases of conflict.[9]

The cells of the matrix indicate the players' expected payoffs from each strategy profile.[10] The payoffs for the three coordination outcomes can be represented quite simply. We let $V > 0$ denote the total amount of resources (the "gains from cooperation") to be divided between the two players in each instance of conflict. Since coordinating on ROW gives Rowan everything in every period, Rowan gets V and Collins gets 0 if they

8 To simplify our exposition, we restrict our attention to a two-player game. We could have presented an n-player model, which produces qualitatively similar conclusions, but the added complexity would make our argument less intuitive. Additionally, although Moehler claims that the empathetic contractor model permits n players (112), his analysis often proceeds by way of two-person cases (e.g., 124).

9 We address Moehler's restriction to these three strategies in the next section.

10 To simplify the exposition, we assume that each player's utility function is linear in their share of resources; formally, for $i \in \{R, C\}$, $u_i(x) = x$, where x is i's share of resources.

coordinate on ROW; similarly, Collins gets V and Rowan gets 0 if they coordinate on COL. For now, we let $W_R(V)$ and $W_C(V)$ denote each player's share of V if they coordinate on the WPU. We assume that each player's share from coordinating on the WPU is less than everything and more than nothing; that is, $V > W_i(V) > 0$ for $i \in \{R, C\}$. Since $W_i(V)$ is determined in part by the players' relative bargaining capacities (in accordance with the WPU), and since these might be unequal, we do not assume that $W_R(V) = W_C(V)$.

We now need to specify the players' expected payoffs from failing to coordinate on any principle. According to Moehler, failure to coordinate results in violent conflict (126; also 19, 27, 98–9). In setting the payoffs for violent conflict, we must recognize that conflict is costly in the sense that it wastes resources. We can represent this using a cost parameter $c > 0$ and say that, upon the completion of hostilities, $V - c$ is what's left for the players to divide amongst themselves. How exactly is $V - c$ divided? To cohere with the fact that the peace game is supposed to model a repeated interaction in which violence in each period is a possible outcome, we assume that the players fight until they reach a stalemate, in which case they are each left with a percentage of $V - c$. We let p_R and p_C denote the two players' stalemate shares of $V - c$, with p_R being a number in the interval $[0, 1]$ and $p_C = 1 - p_R$. For simplicity, we label each player's payoff from a single conflict as $d_R = p_R(V - c)$ and $d_C = p_C(V - c)$ ("d" for "disagreement"). We leave open the possibility that one player is strong enough to take everything in a fight; formally, we allow that $d_i = 0$ for some player i.

In sum, we believe Moehler's peace game is best represented as an indefinitely repeated version of the stage game represented in figure 1. Before turning to Moehler's claim that coordinating on the WPU is the unique mutually rational solution, we specify the appropriate notion of a mutually rational solution. In game theory, the intuitive notion of a mutually rational outcome is captured by the general concept of an equilibrium strategy profile. Roughly, an equilibrium characterizes a strategy profile in which each player is doing the best she can given her preferences over outcomes and her expectations about others' behavior. The appropriate equilibrium concept for the kind of repeated game Moehler has in mind is *subgame perfect equilibrium*. Hence, we read Moehler's claim that coordination on the WPU is the uniquely rational outcome as equivalent to the claim that it is the unique subgame perfect equilibrium of the game. A *strategy* specifies which action to perform in each stage game (i.e., instance of conflict), sometimes allowing this to depend on what has happened in past stages. To say, therefore, that coordination on the WPU is the uniquely rational strategy is to say that the sole mutually rational strategy for each player is to comply with the WPU at each stage.

Moehler simplifies the task of identifying mutually rational strategy profiles by re-

stricting his attention to strategies that instruct players to perform the same action at each stage of the interaction regardless of which actions were performed at past stages (122–23).[11] This means that both players are restricted to the following strategies:

- Unconditional WPU: *divide V according to the WPU in each case of conflict regardless of how past conflicts have been resolved*;
- Unconditional ROW: *give Rowan everything in each case of conflict regardless of how past conflicts have been resolved.*
- Unconditional COL: *give Collins everything in each case of conflict regardless of how past conflicts have been resolved.*

Moehler's rationale for this restriction is that only these types of strategies can "ensure peaceful long-term cooperation" (126). We criticize this restriction in the next section.

Moehler's restriction reduces an infinite number of admissible strategy profiles for the repeated game to the nine profiles corresponding to the cells of the stage game in figure 1. This means that we can use the matrix for the stage game to represent the repeated game with limited strategy options. To simplify our exposition, we assume that V, d_i, and $W_i(V)$ are the same in each period; hence, we treat each cell as reporting the per-period payoffs for the stage game, from which we can readily calculate the expected payoffs for unconditionally repeating the actions corresponding to each cell.[12]

We take Unconditional WPU to specify the strategy Moehler has in mind when he says that coordinating on the WPU is the uniquely rational outcome of the peace game. We must now ask: (1) whether a strategy profile for which both players play Unconditional WPU is a subgame perfect equilibrium for the repeated game, and (2) whether this profile represents the only such equilibrium.

11 In game-theoretic terms, Moehler restricts his attention to "stationary", "pure", and "unconditional" strategies. Note that Moehler conflates the mixed strategy and conditional strategy concepts (126).

12 The expected payoffs for the entire game are a discounted sum of the payoffs in each period. For example, Rowan's expected payoff (as judged from the present time period) when both players play Unconditional WPU is

$$W_R(V) + \delta W_R(V) + \delta^2 W_R(V) + \ldots + \delta^\infty W_R(V),$$

where $0 \leq \delta \leq 1$ is the factor by which payoffs in future periods are discounted. This sum simplifies to $\frac{W_R(V)}{1-\delta}$. Similarly, Rowan's expected payoff from playing Unconditional ROW when Collins plays Unconditional COL simplifies to $\frac{d_R}{1-\delta}$. Since $\frac{1}{1-\delta}$ is a common scalar for both payoffs, we can multiply both by $(1-\delta)$, which leaves us comparing $W_R(V)$ and d_R. But these are just the per-period payoffs reported for the corresponding cells in figure 1. Our analysis can be readily generalized to the case where the per-period payoffs corresponding to each cell are allowed to vary.

Our analysis here requires specifying the relative value of the expected payoffs. To start, we assume that getting all of the available resources every time is better than getting nothing every time ($V > 0$), and it is also better than the expected payoff from indefinite violent conflict ($V > d_i$ for $i \in \{R, C\}$). We also assume that the expected value of indefinitely resolving conflict in accordance with the WPU is greater than the expected value of settling every conflict with violence ($W_i(V) > d_i$ for $i \in \{R, C\}$). Assuming otherwise would imply that the players have no incentive to resolve conflicts without violence. The last assumption is sufficient to vindicate (1): coordinating on the WPU is an equilibrium.[13] So far, so good.

To assess (2) whether coordinating on the WPU is the *unique* equilibrium, let's adopt Rowan's perspective (without loss of generality) and suppose that Collins insists on taking everything all the time, that is, Collins plays Unconditional COL. What should Rowan play? If he plays Unconditional ROW or Unconditional WPU, his payoff at any stage is d_R; if he plays Unconditional COL — if he accedes to Collins's demand every time — then his payoff is always 0. When specifying the payoffs above, we assumed that $d_R \geq 0$. Given this, and continuing with our assumption that Collins plays Unconditional COL, Rowan has two equal-best responses: play Unconditional ROW or Unconditional WPU.[14] Thus, if $d_R \geq 0$, there is a *conflict equilibrium* in which both insist on taking everything all the time (i.e., Rowan plays Unconditional ROW and Collins plays Unconditional COL). In other words, coordinating on the WPU is not the unique equilibrium for the peace game if $d_R \geq 0$.

Moehler has two potential routes to excluding the conflict equilibrium. First, he suggests at places that *homo prudens* prefers cooperation above all else (e.g., 27, 87, 101-102, 104). This might suggest what is contrary to our assumption above, namely, that Rowan would avoid conflict with an intransigent Collins even if it means taking a loss (and *vice versa*) (i.e., $d_i < 0$). But if we allow this possibility, then coordinating on COL and coordinating on ROW are also equilibria. So this route excludes the conflict equilibrium but still leads to multiple equilibria, contrary to Moehler's uniqueness claim.[15]

The second route for excluding the conflict equilibrium requires returning to Rowan's

13 We relegate all proofs of this and following statements about the model's equilibria to a technical appendix.

14 Coordinating on COL is also an equilibrium if $d_R = 0$, which is shown by reasoning analogous to the proof for proposition 3 in the appendix.

15 In fact, Moehler seems to qualify this claim about the players' preferences, suggesting that they would prefer conflict over simply conceding everything to the other player, their strong preference for cooperation notwithstanding (122–23). Thus, it seems more charitable to interpret Moehler as assuming that $d_i \geq 0$ for $i \in \{R, C\}$, as we did above.

	COL	WPU
ROW	d_R, d_C	d_R, d_C
WPU	d_R, d_C	$W_R(V), W_C(V)$

Figure 2. Reduced "peace game"

perspective and assuming that $d_R > 0$. Now observe the following: if Collins plays Unconditional ROW, then Rowan's best response is Unconditional ROW; if Collins plays Unconditional COL, then Rowan's best response is either Unconditional ROW or Unconditional WPU; if Collins plays Unconditional WPU, then Rowan's best response is Unconditional WPU. The crucial insight here is that, whatever Collins's strategy, Rowan never plays Unconditional COL (in game-theoretic terms, Unconditional COL is not *rationalizable* for Rowan). Collins knows this — that is, Collins knows that Rowan is playing either Unconditional ROW or Unconditional WPU.[16] By similar reasoning, Collins never plays Unconditional ROW and Rowan knows that Collins is playing either Unconditional COL or Unconditional WPU. To represent the players' epistemic situation, we can cross out the COL row and ROW column in figure 1, which leaves us with the reduced game in figure 2. Knowing that Collins conforms to the *homo prudens* model of rationality, Rowan can anticipate Collins's view of the situation: "If I play Unconditional COL, we're guaranteed to wind up in indefinite violent conflict; if I play Unconditional WPU instead, there's a chance that we end up in indefinite violent conflict, but there's also a chance that we coordinate on the WPU. Given my strong preference for avoiding violence, I should at least give myself a chance of peaceful cooperation" (cf. 122–123). (In game-theoretic language, playing COL/ROW is *weakly dominated* by WPU.) In view of this reasoning, it seems unreasonable for Rowan to expect Collins to play Unconditional COL. Given this expectation, it is not rational for Rowan to play Unconditional ROW. While this line of reasoning does not strictly imply that *both insist on taking everything* is not an equilibrium, it plausibly demonstrates that it can be ruled out by reasonable expectations about how *homo prudens* would choose in this situation.

Our analysis in this section implies that, strictly speaking, there are multiple equilibria in the peace game. Still, a plausible line of reasoning about how *homo prudens* views her choice in the peace game can be used to rule out all equilibria aside from coordinating on the WPU. In the next section, we use our model to expose some of the strong assumptions

16 Collins knows this because standard game-theoretic models assume that the rationality of the players is common knowledge.

that are crucial to delivering this result and consider how Moehler might justify these assumptions.

3. IS THE WPU UNIQUELY RATIONAL?

In this section, we focus on the assumptions underlying Moehler's exclusion of various plausible candidate principles as strategies. We show that, once these restrictions are lifted, coordinating on the WPU is not uniquely rational. Moreover, we show that the justification Moehler might offer for these restrictions reveals that his argument deviates in unacknowledged ways from the standard (Bayesian) conception of instrumental rationality.

Before we scrutinize the restrictions Moehler uses to secure his uniqueness result, we briefly explain why they are noteworthy from the perspective of his chosen analytical framework, namely, the theory of repeated games. The key point of interest is that, by restricting our attention to the three unconditional strategies specified above, Moehler removes from consideration a vast number of mutually rational strategy profiles, each of which produces an outcome that is Pareto superior to indefinite violent conflict. We can provide some intuition for this point using figure 3. The horizontal axis measures Rowan's average per-period payoff as delivered by some specified strategy profile and the vertical axis likewise measures Collins's average per-period payoff; the dashed lines indicate the players' average per-period payoffs from indefinite conflict. Moehler's restrictions leave nine possible strategy profiles for the "peace game", one for each pair of actions available in the stage game (recall fig. 1). Put in terms of the players' per-period payoffs (reported in the cells of fig. 1), Moehler's restrictions on the players' strategies imply that the "peace game" has four possible outcomes, which are indicated by the four labeled points in figure 3: $(0,V),(V,0),(d_R,d_C),(W_R(V),W_C(V))$.[17] The figure makes clear that, given Moehler's restrictions, coordinating on the WPU is the only outcome that is Pareto superior to indefinite conflict.

Without Moehler's restrictions, however, the theory of repeated games implies that the "peace game" has infinitely many possible payoff outcomes, one for each point in

17 Note that the players get the same average conflict payoffs for six of the nine possible outcomes. For simplicity, we assume here that $d_R = d_C$, but it is important for Moehler's purposes that this need not be the case. We relax this assumption below. We calculate the WPU payoffs here using the Nash bargaining solution. Moehler calls the WPU a "generalized and universalized statement of the stabilized Nash bargaining solution" (127), the latter of which is Moehler's amended version of the standard Nash solution. The stabilized and standard Nash solutions deliver identical payoffs in many cases. We consider a case where they come apart below.

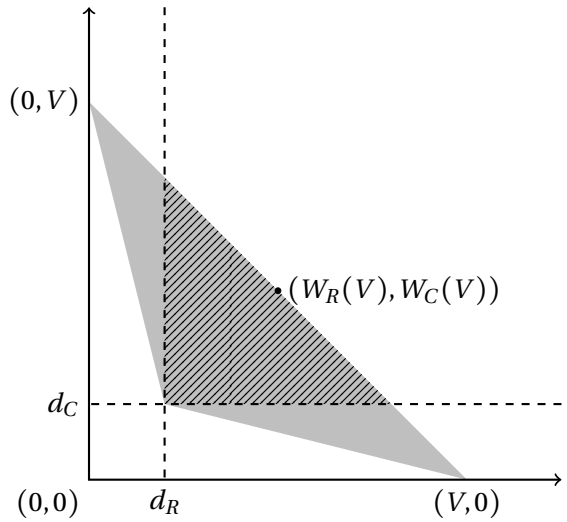

Figure 3. Payoff possibilities for repeating the game in fig. 1

the light grey triangle in figure 3.[18] More strikingly for our purposes, the so-called "folk theorem" — a central result in the theory of repeated games — implies that every point in the hatched region to the northeast of the dashed lines corresponds to the average per-period payoffs for some equilibrium strategy profile for the repeated game.[19] Notice that every point in this hatched region corresponds to a distribution of V that is Pareto superior to the distribution that results from violent conflict; the set of Pareto optimal distributions lies on the northeast edge of the hatched triangle. Since there are infinitely many points in the indicated region, the folk theorem implies that there are infinitely many equilibria for the repeated game, and each of these yields per-period average payoffs that are Pareto superior to violent conflict. In light of this, how might Moehler justify the restrictions he imposes on the strategies available in the "peace game"?

To interrogate Moehler's reasoning here, we limit our attention to a class of principles that prescribe a simple division of V in each period. Let a *β-principle* prescribe the division $\beta V, (1-\beta)V$; for instance, the $\frac{3}{8}$-principle ($\beta = \frac{3}{8}$) prescribes that, in each period, Rowan gets $\frac{3}{8}V$ while Collins gets $\frac{5}{8}V$. There are infinitely many β-principles that prescribe

18 The light grey triangle represents the "convex hull" of the possible outcomes for the stage game, i.e., the set of all possible probability mixtures over the payoff possibilities reported in fig. 1.

19 Slightly more precisely: for every point (x, y) in the grey triangle such that $x > d_R$ and $y > d_C$, there is an equilibrium strategy profile for which the average per-period payoffs are arbitrarily close to (x, y). See Fudenberg and Tirole (1991, 150–60, 192–97).

a Pareto optimal division of V, one for each $0 < \beta < 1$ such that $\beta V \geq d_R$ and $(1-\beta)V \geq d_C$. We can show that, if Pareto optimal β-principles are included among the available strategies for the "peace game", then coordinating on the WPU is no longer the unique mutually rational strategy. For any β-principle such that $\beta V > d_R$ and $(1-\beta)V > d_C$, neither player has any incentive to deviate from the specified division at any stage.[20] Including Pareto optimal β-principles thus brings infinitely many coordination equilibria. Moreover, coordinating on these principles is mutually beneficial (relative to violent conflict) and, provided the players can coordinate, these principles can be easily institutionalized (formally or informally) since they give clear and determinate instructions.

What, then, can justify excluding β-principles?[21] Importantly for Moehler's purposes, β-principles prescribe a division of V without making any claim about the moral rectitude of the resultant division; so he cannot exclude them on the grounds that they are problematically moralized. Although he never explicitly considers β-principles, Moehler's discussion of rational agents' "fundamental interests" presents potential reasons against their inclusion.

According to Moehler, rational agents have two "fundamental interests", which he characterizes as requirements of instrumental rationality. The first is that, in cases of conflict, agents must be allowed "to defend their interests maximally based on their actual capacities" (115). The second, which we set aside for discussion in the next section, stipulates an interest in resolving conflict from a position in which one's basic needs are secured. Moehler argues that the first interest supports adoption of the classic Nash bargaining solution (NBS) as a default principle of conflict resolution. The NBS assigns to each player her disagreement share plus some non-negative amount of any remaining cooperative surplus. Insofar as a player has a superior starting position, she can secure for herself a higher share of resources. Moehler takes it that this models agents' capacity to maximize their utility subject only to "ethically neutral" constraints on their bargaining capacity (55–6, 85).

Given Moehler's reasons for taking the NBS as a default principle, we conjecture that he would exclude β-principles on the grounds that they are insensitive to agents' actual bargaining capacities.[22] To illustrate, imagine Rowan and Collins are considering whether to coordinate on the $\frac{3}{8}$-principle, which gives them (respectively) $\frac{3}{8}V$ and $\frac{5}{8}V$ in each period. Suppose, in contrast, that dividing V according to their relative bargaining

20 The reasoning for this claim is analogous to the reasoning in the proof for proposition 2 in the appendix.

21 Perhaps one argument for excluding them is that one might think it impossible to coordinate on a single equilibrium out of an infinity of options. But that doesn't entail that coordination on any of the infinite equilibria is *irrational*; it only questions the possibility of coordinating on a specific equilibrium.

22 Moehler rejects Gauthier's proposal on precisely these grounds (55–6).

	$\left(\frac{3}{8}, \frac{5}{8}\right)$	NBS	$\cdots$
$\left(\frac{3}{8}, \frac{5}{8}\right)$	$\frac{3}{8}V, \frac{5}{8}V$	d_R, d_C	$\cdots$
NBS	d_R, d_C	$\frac{3}{4}V, \frac{1}{4}V$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\ddots$

Figure 4. Peace game with β-principles

capacities would instead give Rowan $\frac{3}{4}V$ and leave Collins with the remainder (see figure 4). Following Moehler's reasoning, it might appear irrational for Rowan to agree to the principle that gives him less than he could expect from a principle that divides V according to relative bargaining power (e.g., the NBS). But this reasoning presupposes that Rowan is in a position to agree to one principle or another; this, in turn, presupposes that the two players are bargaining over candidate principles. Moehler, however, provides no bargaining protocol for the peace game: there is no provision for the players to propose a principle for adoption and, so, no provision for players to accept or reject proposals (108). In the peace game, what is rational for Rowan to do depends on his beliefs about which strategy Collins is playing (and vice versa). As shown in figure 4, if Rowan believes Collins is playing the $\frac{3}{8}$-principle and Rowan's share from coordinating on this principle is greater than his violent conflict share ($\frac{3}{8}V > d_R$), then it is rational for Rowan to coordinate on this principle — even if he would prefer to coordinate on the NBS. This is a general feature of noncooperative game theory: because players' strategies interdependently determine outcomes, rational behavior is not simply a matter of unilaterally adopting the strategy associated with one's most favored outcome.

To consolidate the previous point, consider the scenario in figure 4 from Collins's perspective. Suppose Rowan is playing NBS and suppose $\frac{1}{4}V > d_C$. In this case, we might be less tempted to say that Collins would behave irrationally by coordinating on the NBS because she could do better if the players were to coordinate instead on the $\frac{3}{8}$-principle. Indeed, we might be tempted to say that it is rational for Collins to coordinate on the NBS because $\frac{1}{4}V$ is "the best she can expect to get in light of her lesser bargaining power". But this reasoning conflates two distinct issues: that of identifying which strategy profiles are mutually rational, and that of selecting among multiple mutually rational profiles. It is rational for Collins to coordinate on NBS *if and only if Rowan is playing NBS*, whatever the players' relative bargaining power. Facts about relative bargaining power can help players select among multiple mutually rational strategy profiles — if those facts are common

knowledge among the players and if those facts are salient and thus help focus players' joint attention on a particular strategy profile.[23] But facts about relative bargaining power do not settle *which strategy profiles are mutually rational*. Equilibrium identification is distinct from equilibrium selection.

To sum up: Moehler argues that rational agents would choose to resolve conflicts using a bargaining principle such as the NBS or the WPU because instrumental rationality requires that they choose a principle that allows them "to defend their interests maximally based on their actual capacities" (10, 85, 101, 119–20). But this reasoning mistakes an outcome of instrumentally rational behavior in a specific context for a general requirement of rationality. Most generally, the standard theory of instrumental rationality requires that agents maximize their expected utility subject to whatever constraints they face in pursuing their ends. In the context of noncooperative games, the key constraints — aside from one's own strategy options — are other players' strategy choices. As defined by the Nash equilibrium concept and its refinements, instrumentally rational agents stuck in noncooperative strategic situations maximize their expected utility subject to their beliefs about the other players' strategies. With respect to noncooperative bargaining in particular, there is no requirement that players *enter the bargaining situation resolved* to divide resources according to their relative bargaining capacities. Indeed, a noncooperative strategic situation is defined as a situation in which players are unable to commit to playing a particular strategy prior to entering the interaction. There is thus no standard rationality concept for noncooperative games that *defines* a mutually rational strategy profile in terms of players' relative bargaining power.[24]

We have shown that Moehler's first "fundamental interest" is not a dictate of what he call "orthodox instrumental rationality".[25] Thus, it cannot be used to exclude nonmoralized Pareto optimal β-principles from the set of candidate principles absent an argument to show that this constraint is part of a broader conception of rational prudence that contractarians can accept. We have yet to find an argument to this effect.

Further, we have shown that a proper application of the standard view of instrumental rationality to the peace game implies that it is mutually rational for agents to coordinate on any of an infinite number of β-principles. The key lesson of our argument is that

23 Schelling (1960) is the classic source on the use of "focal points" to select among multiple equilibria.

24 It is true that a bargaining game can result in an equilibrium division of V that could be interpreted as sensitive to players' relative bargaining capacities (e.g., Binmore, 1998; Rubinstein, 1982). But this is not because players enter the bargaining situation under a requirement to divide V in this way. It is instead a *result* of agents' efforts to maximize expected utility subject to their beliefs about others' strategies.

25 Note well that Moehler follows others in criticizing Gauthier's contractarian theory for adopting a "nonorthodox concept of rationality" (51, 56–7).

Moehler's game-theoretic argument for the WPU is inconsistent with his ambition to derive a unique principle of conflict resolution from nothing but the standard conception of instrumental rationality. Moehler thus faces a choice. If he persists in thinking about cases of conflict as a noncooperative coordination game, he can opt for the original model in section 2 or the amended model considered in this section. In the former case, coordinating on the WPU is uniquely "rational" but only if we adopt a non-standard requirement of instrumental rationality, one that constrains players to adopt strategies that divide resources according to their relative bargaining capacities. If Moehler eschews this non-standard restriction on agents' coordination possibilities, then instrumental rationality is consistent with coordination on numerous principles of conflict resolution.[26]

4. IS STABILIZATION INSTRUMENTALLY RATIONAL?

Our argument thus far leaves it open to Moehler to argue for a weaker claim, namely, that rational agents have strictly prudential reasons to choose the WPU over the standard Nash bargaining solution (see 84–92 and 127–29). Although this claim is less striking than Moehler's claim that the WPU is the uniquely rational principle of conflict resolution, it should nonetheless hold considerable interest for contractarians, who typically justify their principles as favored alternatives to the Nash solution (e.g., Gauthier 1986; for more recent discussion, see Thrasher 2013 and Thoma 2015). Moehler's argument for this weaker claim puts him squarely within this contractarian tradition and, if successful, would constitute an argument for a novel alternative to the Nash solution. In this section, we show that positing a second "fundamental interest" is necessary to vindicate this argument. Unfortunately, like the first fundamental interest, this too imports normative resources beyond those contained in the standard theory of instrumental rationality. The upshot is that instrumentally rational agents (conventionally understood) have no incentive to adopt the WPU over the NBS.

Moehler's argument for the claim that the WPU is rationally superior to the NBS depends on concerns about the long-term stability of peaceful cooperation (127–28). For

26 Our critique in this section has granted Moehler's claim that traditional moral principles cannot serve to coordinate rational agents' behavior in circumstances of conflict. Note, however, that this claim seems inconsistent with Moehler's view of *homo prudens*. To wit, one might think that the WPU states the moral truth. If so, then disagreement among agents about whether the WPU or (e.g.) the principle of average utility is morally correct might result in conflict. Yet Moehler clearly thinks that rational agents can abstract from this moral disagreement and coordinate on the WPU for solely prudential reasons (see 124). But then, they should be able to do the same for other "morally thick" principles; at any rate, Moehler gives no reason to think that they cannot do so.

Moehler, the NBS "represents a natural agreement point for rational agents" in one-off conflict situations (2018, 56; see also Moehler 2010, 447, 453). Once long-term considerations are brought to bear on the choice of principles, however, Moehler argues that rational agents would reject the NBS because agents who coordinate on the NBS are not guaranteed to enter or leave negotiations from a position where their basic needs are satisfied. According to Moehler, agents who are left with an inadequate share of resources have an incentive to use violence to take what they can, which undermines peaceful cooperation (115–16). The NBS is thus unsatisfactory because cooperation based on it is unstable. The WPU is supposed to stabilize the NBS by guaranteeing that agents be allowed to resolve their conflicts from a position where their basic needs are met. So understood, Moehler's argument for the WPU has two steps. First, he argues that, absent concerns about long-term stability, rational agents in the relevant choice situation would coordinate on the NBS. Second, he argues that agents cannot ignore stability considerations and that, once these concerns are introduced, agents would coordinate on a stabilized version of the NBS, which is just the WPU. We can call this Moehler's "stabilization argument" for the WPU.

We note that the WPU is naturally thought of as being constructed "on top of" the NBS in a certain way: the WPU prescribes that agents use the standard NBS to divide resources among themselves, but only after a "stabilizing adjustment" has been made to agents' expected payoffs from conflict (if necessary) to ensure that each agent begins the conflict resolution process from a position in which their basic needs are guaranteed to be met. We will need to get clear on how to interpret this "stabilization constraint" shortly, but for now, the important point to notice is that the NBS and the WPU will sometimes prescribe the same division of resources. In at least some cases of conflict, each party's "disagreement payoff" (i.e., their prospects given conflict) will be sufficient to ensure that their basic needs will be met should cooperation break down and conflict ensue; in these cases, the WPU simply prescribes the standard NBS division of resources. Where the WPU prescribes a division that departs from the standard NBS, it is because the WPU requires that a stabilizing adjustment first be made to at least one agent's disagreement payoff to ensure that each agent has "at least the means that they need to maintain their existence as separate agents and satisfy their basic needs as a basis for conflict resolution" (115). Once this adjustment has been made, the WPU applies the NBS using these adjusted disagreement payoffs as the basis for dividing the remaining resources.

A key feature of both the NBS and the WPU is that they prescribe divisions of the gains from cooperation (i.e., V) as a function of the players' disagreement payoffs. Moehler follows convention in interpreting these disagreement payoffs as indicating Rowan's and

Collins's respective bargaining strength. For instance, if both Rowan and Collins expect that Rowan will receive a higher payoff than Collins should violent conflict break out, then it is plausible to think that Rowan is better able to push for a peaceful resolution that gives him a greater share of V. Put differently, Rowan's higher expected conflict payoff implies that a higher prescribed cooperation payoff is required to make it rational for Rowan to choose cooperation over conflict. Given this, it is natural to think that Rowan has superior bargaining strength relative to Collins. This interpretation aligns with Moehler's claim that the NBS and WPU are examples of principles that satisfy agents' first "fundamental interest", namely, that conflict be resolved in a manner that is sensitive to the players' relative bargaining strength.

Before we can assess Moehler's claim that it is instrumentally rational to adopt the WPU over the NBS, we must determine the specific division of V each principle prescribes. We focus on the cases where the two principles prescribe different divisions. Recall that the two principles diverge only in those cases where a "stabilizing adjustment" must be made to ensure that each agent's basic needs are met as a basis for conflict resolution. We can now make this idea more concrete: if an agent i's disagreement payoff is insufficient to meet i's basic needs, then a stabilizing adjustment must be credited to i's disagreement payoff prior to calculating the division of V that is to serve as the basis for peaceful cooperation. Let d^* denote the minimum share of goods agents "need to maintain their existence as separate agents and satisfy their basic needs" (115). Then we can say that Rowan's disagreement payoff is insufficient to meet his basic needs if $d_R < d^*$, and similarly for Collins if $d_C < d^*$. Without loss of generality, we will assume that $d_C < d^* < d_R$; this implies, first, that Rowan has greater bargaining strength than Collins and, second, that a "stabilizing adjustment" must be made to ensure that Collins's basic needs are met before deciding how to divide V.

The standard NBS prescribes that the players divide V so as to maximize the Nash function:

$$(x_R - d_R)(V - x_R - d_C), \tag{1}$$

where x_R is Rowan's share and Collins gets the remainder, $x_C = V - x_R$. The resultant division gives Rowan and Collins the following payoffs:

$$x_R^N = \frac{1}{2}(V - d_C + d_R) \qquad\qquad x_C^N = V - x_R^N = \frac{1}{2}(V - d_R + d_C). \tag{2}$$

This case is depicted in figure 5(a) using the solid lines that cross at the point (x_R^N, x_C^N), which indicates the NBS division. Since, by hypothesis, Rowan is in a stronger bargaining

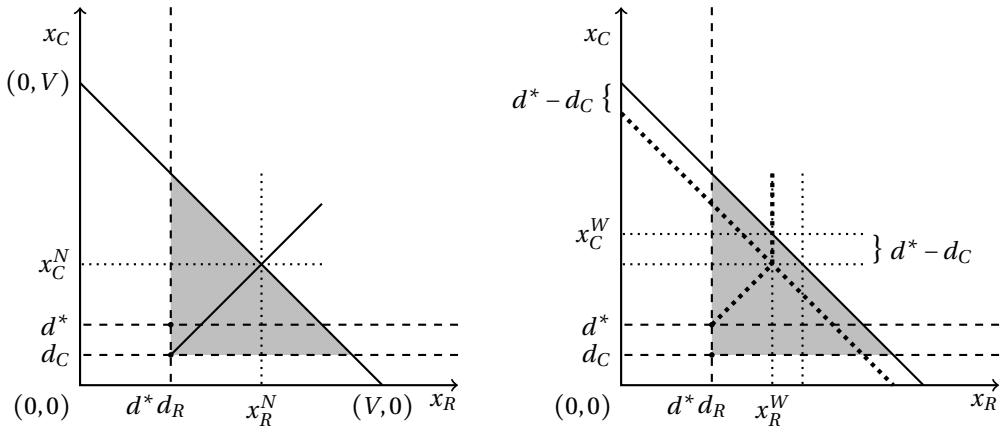

Figure 5

position than Collins, the NBS gives Rowan a higher share, in accordance with Moehler's first "fundamental interest".[27]

In contrast with the NBS, the WPU requires that Collins's disagreement point be adjusted upward from d_C to d^* prior to calculating the players' shares of V. One immediate question is where the resources are to come from to make up the shortfall $d^* - d_C$. We assume they come out of V, for two reasons. First, if the WPU were to require transfers from Rowan to Collins, then, in cases where $d^* - d_C$ is greater than $d_R - d^*$, bringing Collins up to d^* would require dropping Rowan below that same threshold. Second, Moehler repeatedly says that basic needs are to be satisfied prior to bargaining "if the goods that are in dispute [i.e., V] permit it" (e.g., 118). This suggests that, prior to dividing V, an amount $d^* - d_C$ is to be credited to Collins from V if possible; following this, the remaining resources $V - (d^* - d_C)$ are to be divided between the two players according to the NBS. Absent guidance to the contrary, we will assume that this is what the WPU requires.

The implications of the WPU adjustments are depicted in figure 5(b), with the negative sloping dotted line indicating the frontier of the bargaining space after the stabilizing transfer of $d^* - d_C$ has been credited to Collins. Applying the NBS to this new situation (as the WPU requires), the players are to divide what remains after the stabilizing adjustment,

27 Notice that, if $d_R = d_C$, then each player gets $\frac{1}{2}V$.

$V - (d^* - d_C) = V - d^* + d_C$, so as to maximize the adjusted Nash function, assuming Collins's disagreement payoff is d^* rather than d_C:

$$(x_R - d_R)([V - d^* + d_C] - x_R - d^*), \tag{3}$$

where x_R again denotes Rowan's share; Collins again gets the remainder, $x_C = V - x_R$ (that is, Collins gets $[V - d^* + d_C] - x_R$ plus the stabilizing credit $d^* - d_C$). The resultant division gives Rowan and Collins the following payoffs:

$$\begin{aligned} x_R^W &= \frac{1}{2}(V - d_C + d_R) - (d^* - d_C) \\ x_C^W &= V - x_R^W = \frac{1}{2}(V - d_R + d_C) + (d^* - d_C). \end{aligned} \tag{4}$$

Having determined the divisions prescribed by the two principles, we turn to Moehler's claim that rational agents would adopt the WPU over the NBS. To facilitate our comparison, notice that we can rewrite the players' WPU payoffs as a function of their NBS payoffs:

$$x_R^W = x_R^N - (d^* - d_C) \qquad\qquad x_C^W = x_C^N + (d^* - d_C). \tag{5}$$

We can see immediately that, in cases where the NBS and WPU make different prescriptions, the players have conflicting preferences over the two principles: Rowan strictly prefers the NBS and Collins strictly prefers the WPU (since $d^* - d_c > 0$ by assumption).[28] Moehler must show that both players nonetheless have a rational incentive to adopt the WPU rather than the NBS.

According to Moehler's stabilization argument, the adjustment prescribed by the WPU is a requirement of instrumental rationality: players who do not have sufficient resources to satisfy their basic needs "will take whatever they need in order to survive", which brings violent conflict (118). To avoid conflict, then, it is rational for each player to ensure that all players have enough to satisfy their basic needs. In particular, it is rational for the *stronger* player (i.e., Rowan) to ensure that the weaker player (i.e., Collins) has no incentive to initiate violence in an attempt to satisfy her basic needs. Thus, it is rational to adopt a principle that guarantees each player an adequate minimum share.

Notice, however, that in many cases where the NBS and WPU come apart (e.g., in cases where $d_R > d^* > d_C$, per our assumption), the NBS provides Collins with an adequate

28 Notice also that Rowan ends up paying for the stabilizing adjustment in the form of foregone resources (relative to his NBS payoff) even though we assumed that $d^* - d_C$ was taken from V prior to bargaining.

minimum share of goods even without the WPU's stabilizing adjustment; figure 5(a) presents one such case, while also indicating how numerous other such cases could be constructed. We submit that, in these cases, Moehler's argument is unmotivated: standard Nash bargaining would meet Collins's basic needs even without the adjustment to her disagreement payoff. Collins would, of course, prefer the WPU to the NBS, but that's not enough to show that rational prudence requires *both* players to adopt the WPU over the NBS, especially given that Rowan has conflicting preferences. And since the NBS provides Collins with a minimally adequate share of resources, she has no incentive (per Moehler's *homo prudens* model) to initiate violence by insisting on the WPU.[29] Rowan, in turn, has no rational incentive to transfer resources to Collins as a means to avoid violent conflict.

Moehler would reply here that we have ignored rational agents' second "fundamental interest", namely, that rational prudence requires agents to adopt a principle that guarantees a minimally adequate share of resources "*as a basis for conflict resolution*" (115, emphasis added). As he puts it, "it would be irrational for agents to select a principle of conflict resolution that does not secure their minimum standards of living. . . *as a starting point for conflict resolution*" (118, emphasis added). If this is right, then it is rational for both players to adopt the WPU over the NBS since only the WPU guarantees both players a minimally adequate share "as a starting point for conflict resolution" (see also 86–9). Hence, the rationality of the WPU depends on this second "fundamental interest", which our argument above neglects.

This, however, marks a subtle departure from the standard theory of instrumental rationality. As we noted in section 3, the standard theory requires only that rational agents maximize their expected payoffs subject to whatever constraints they face in pursuing their ends. There is no *general* requirement of rationality that agents start their noncooperative conflict resolution process from a position that guarantees both players an adequate minimum share of resources. Of course, both players could find it rational to guarantee an adequate minimum if, for example, each expects that leaving a player with less than an adequate minimum will decrease their respective expected payoffs over the long run. But if so, then the rationality of guaranteeing an adequate minimum is not imposed as an ex ante requirement but instead *emerges endogenously* from the features of a particular type of strategic interaction. We've already seen that there are some cases where the players have no reason to expect adverse consequences from using the NBS rather than the WPU: when the NBS gives Collins a minimally adequate share without the stabilizing adjustment, she has no rational incentive to initiate conflict; so Rowan has no

29 Recall Moehler's claim, noted above, that agents who insist on resolving conflict in accord with their most preferred principle at the cost of violent conflict do not fit the *homo prudens* model.

rational incentive to make the transfer.

Moehler could reply that a rational incentive to guarantee an adequate minimum does emerge endogenously because resolving conflict using the NBS risks leaving some players with less than an acceptable minimum, which leaves them vulnerable to an unacceptable fate. Intuitively, weaker agents who are left without adequate resources will find it rational to use violence in an effort to avoid this worst-case scenario. Stronger agents faced with this threat of violent conflict will, in turn, find it rational to provide weaker agents with an adequate minimum in an effort to preempt costly violent conflict (118–20, 135, 168). If so, then even if there is no general requirement that conflict resolution start from a position that guarantees an adequate minimum, the risks inherent in their situation will lead both players to find it rational to adopt the WPU over the NBS to insure against the violent conflict they both seek to avoid.

However intuitive this line of reasoning might be, we can readily show that it is invalid. There are two propositions to consider: first, that it is rational for Collins to initiate violent conflict when the NBS leaves her with less than an adequate minimum ($x_C^N < d^*$) by, for example, insisting on the WPU; second, that it is rational for Rowan to preempt Collins's threat of violence by coordinating on the WPU.

Regarding the first proposition, notice that, even if Collins's NBS payoff is less than an adequate minimum, it is still greater than her conflict payoff so long as there are sufficient resources to rationalize cooperation:

$$
\begin{aligned}
V &> d_R + d_C && \text{(Mutually rational cooperation is possible)} \\
V - d_R + d_C &> 2d_C \\
\frac{1}{2}(V - d_R + d_C) &> d_C \\
x_C^N &> d_C && \text{(By definition of } x_C^N\text{)}
\end{aligned}
$$

This ostensibly indicates that Collins has no rational incentive to initiate conflict even if her NBS payoff is less than an adequate minimum. But we shouldn't accept this conclusion too quickly. Suppose that Collins insists on resolving conflict in accord with the WPU; this is effectively a threat to initiate conflict should Rowan adopt any other principle. If this threat could induce Rowan to coordinate on the WPU, then it may be rational for Collins to threaten violence even if $x_C^N > d_C$. Whether insisting on the WPU is rational, however, depends on whether the threat of violence moves Rowan to abandon the NBS and coordinate on the WPU. For the sake of argument, then, we will follow Moehler in assuming that Collins threatens violence when $x_C^N < d^*$ and consider whether the risk of violence induces Rowan to coordinate on the WPU.

Turning our attention to Rowan, then, we find a striking result: it is not rational for him to abandon the NBS for the WPU in an effort to preempt violence *even if he believes that Collins's threat of violence is credible*. To see this, assume that Collins' share under the NBS is inadequate to meet her basic needs ($x_C^N < d^*$). We can show that this assumption entails that Rowan prefers violent conflict to his WPU share.

$$
\begin{aligned}
x_C^N &< d^* && \text{(Collins threatens violence)} \\
\frac{1}{2}(V - d_R + d_C) &< d^* && \text{(By definition of } x_C^N\text{)} \\
V - d_R + d_C &< 2d^* \\
V - d_R + d_C - 2d^* &< 0 \\
V - d_R + d_C - 2d^* + 2d_R &< 2d_R \\
V + d_R - d_C + 2d_C - 2d^* &< 2d_R \\
\frac{1}{2}(V - d_C + d_R) - (d^* - d_C) &< d_R \\
x_R^W &< d_R && \text{(By definition of } x_R^W\text{)}
\end{aligned}
$$

It follows that Rowan adheres to the NBS even if Collins can credibly threaten violence when $x_C^N < d^*$. To better see the strategic logic behind this result, suppose that the players play a coordination game in which the only strategy options are to adhere to the NBS or adhere to the WPU; conflict ensues if the players do not choose the same principle. If both players adhere to the NBS, then they get their respective NBS payoffs, x_R^N and x_C^N, which leaves Collins with less than an adequate minimum (by assumption). Suppose that Collins insists on using the WPU instead and that she will follow through on this demand. How should Rowan respond? If Rowan adheres to the NBS, the players resolve their conflict with violence, leaving Rowan with d_R. If instead Rowan agrees to coordinate on the WPU in an effort to avoid violence, then Rowan gets x_R^W. We've just shown that $x_R^W < d_R$. Thus, even if Collins would rationally initiate violence when the NBS leaves her with an inadequate share of resources, Rowan has no rational incentive to adopt the WPU with the aim of preempting this violence. So even if there is a risk that circumstances will arise in which Collins rationally threatens violence, it is not rational for Rowan to respond to the risk by adopting the WPU: Rowan is rationally impervious to the risk.

This result is sufficiently counterintuitive to warrant explanation. To see why it holds, notice that (by the definition of x_C^N) we can rewrite our assumption that $x_C^N < d^*$ as

follows:

$$\begin{aligned} \frac{1}{2}(V - d_R + d_C) &< d^* \\ V - (d^* - d_C) &< d_R + d^* \end{aligned} \tag{6}$$

Focus on the latter inequality: The quantity on the left is the amount of resources left to be divided among the two players after implementing the stabilizing adjustment required by the WPU (recall our discussion around expression (3)); the quantity on the right is the amount of resources required to make peaceful cooperation based on the WPU mutually rational. The equivalence in (6) provides a straightforward explanation for our counterintuitive result: if the NBS leaves Collins with less than an adequate minimum, then implementing the stabilizing adjustment required by the WPU leaves a stock of resources that is too small to make cooperation based on the WPU mutually rational. To put the point differently: either Collins's conflict payoff is so low relative to d^* or Rowan's conflict payoff is so high (or both) that providing Collins with an adequate minimum leaves too few resources to make cooperation on the basis of the WPU rational for Rowan. This result arises because the WPU, like the NBS, uses players' unadjusted conflict payoffs to determine when peaceful cooperation is mutually rational (see expression (4) above). To be sure, the WPU makes use of these conflict payoffs in a less direct manner than the NBS does (but see expression (5) above). Yet if we grant Moehler the assumption that Collins rationally pursues violence when her NBS share is less than an adequate minimum, then the WPU, like the NBS, makes the possibility of mutually rational cooperation a function of the players' expected disagreement payoffs.[30]

Summarizing, then: Where the NBS and the WPU coincide, the requirements of instrumental rationality are insufficient to justify choosing the WPU over the NBS. So we must turn our attention to the two kinds of cases where the two principles make

30 Moehler might argue that this result can be avoided so long as ex ante stabilizing concessions on the part of Rowan increase the overall size of V in later periods, leaving both players better off over the long term. We register three concerns about this potential line of response. First, that the stabilizing concessions increase the size of V enough to make Rowan weakly prefer making an immediate concession in favor of long-term gains is a further stipulation that goes beyond anything Moehler explicitly argues. Second, although using concessions to increase the size of V can help to rationalize the WPU, such a stipulation re-raises questions of uniqueness (raised in section 3). Concessions that are more or less demanding than those the WPU requires might have an equal or even greater effect on the size of V. Finally, without explicitly modeling this new constraint, it is difficult to know what we must assume to get the desired conclusion. We are therefore unable to assess in advance whether such assumptions are independently plausible or ad hoc. Since Moehler does not explicitly present this counterargument, we take this additional modeling exercise to be outside the scope of this paper.

different prescriptions. In cases where the NBS leaves the weaker player with an adequate minimum share ($x_C^N \geq d^*$), adopting the WPU to "stabilize" cooperation is rationally unmotivated — neither player has a rational incentive to pursue violence. In cases where the NBS leaves the weaker player with an inadequate share ($x_C^N < d^*$), the stronger player has no rational incentive to adopt the WPU to preempt conflict even if the weaker player can credibly threaten violence. Thus instrumental rationality is insufficient to justify the choice of the WPU over the NBS.

5. CONCLUSION

Moehler argues that the WPU is uniquely consistent with the dictates of rational prudence. In section 2, we constructed a model of Moehler's "empathetic contractor" thought experiment as a means to clarify (on Moehler's behalf) the structure of a valid argument for this claim. In section 3, we showed that Moehler's thought experiment cannot identify a uniquely rational principle without severely restricting the set of coordination possibilities. Our model enables us to show that these restrictions crucially depend on a conception of instrumental rationality that deviates from the standard (Bayesian) view. In section 4, we considered the weaker claim that coordinating on the WPU is rationally superior to coordinating on the NBS. Our model enables us to show that Moehler's argument for this claim relies on a concern for basic needs provision that lies outside the standard theory of instrumental rationality.

Our conclusions might not be surprising, for we have shown that Moehler fails to identify a uniquely rational principle that is also normatively compelling, a problem that is well-known to plague contractarian theories (e.g., Gaus 2011, chap. 2; Sugden 1990). Yet, beyond this, we have demonstrated that this problem arises even if we follow Moehler in limiting the scope of contractarian morality to those cases where it seems most promising. Our argument thus deepens our understanding of the limitations of contractarian moral theory. Its problems lie not in contractarians' immodest ambitions but in the modest resources instrumental rationality can offer to moral theorists. If contractarians wish to justify moral principles on prudential grounds, it is likely that they must incorporate theoretical resources beyond the classical requirements of instrumental rationality.

We've shown that Moehler finds additional resources for his own project in the two "fundamental interests" he attributes to rational agents. He does not hide the crucial role that these fundamental interests play in his argument. Our criticism is thus not that he has been unclear about their importance. Rather, our criticism is that he gives the impression that they follow from what he calls "orthodox instrumental rationality" alone, when in fact they require resources external to it; revealing this is a key virtue of the

modeling exercise we have undertaken. The upshot is that Moehler needs to provide some justification for these constraints. For example, he might justify agents' two fundamental interests by showing that they follow from a commitment to respecting rational agents' autonomy or by appeal to a thin conception of fairness (cf. 115–16, 138). Alternatively, he could provide evidence that contractarians have typically taken such constraints for granted, thus adopting no *further* constraints beyond those orthodox contractarians already accept. Finally, he could claim that agents' "fundamental interests" are reflected in empirical findings about real world agents, and thus reasonable constraints to impose upon a model designed for human rationality in particular.[31] Such moves would allow Moehler to avoid the problem of multiple equilibria (section 3) or provide normative grounds for choosing the WPU over the NBS (section 4). But each of these moves would also compromise the project of justifying the WPU solely by appeal to the demands of instrumental rationality, conventionally understood.

We close by considering an objection that our use of formal game theory is at odds with Moehler's purposes. In justifying his "merely… heuristic" use of game theory, Moehler says that

> despite using basic concepts of game theory for the derivation of the principle of conflict resolution, I do not develop a formal model to derive the principle, because such a model would not necessarily allow me to fully express all of the essential moral considerations of my argument, because the boundaries of moral theory, even for the domain of pure instrumental morality, are not necessarily congruent with the boundaries of game theory, at least not in its standard formulation. (108)

So perhaps we are (uncharitably) imposing on Moehler a burden of proof that he rejects, using methods that block his desired result from the outset. We disagree. Our model in section 2 is a proof of concept: a formal game-theoretic model can indeed be used to reconstruct Moehler's argument for his desired conclusion. In addition, it reveals in an especially clear manner the specific work being done by "the essential moral considerations of [his] argument". More generally, the logic of game theory enables us to

31 At one point, Moehler alludes to this sort of move by claiming that the fundamental interests "follow from the demands of instrumental rationality, basic human needs, and the general empirical conditions under which human beings in this world live" (121). This tacitly admits that the interests require support outside orthodox rationality (in further study of human nature and basic human needs). But it does little to help the reader see how these further resources are being used in support of these particular constraints: He offers no explicit argument.

navigate our way through complex thought experiments so that, by the end, we can be sure that our conclusions are consistent with the standard requirements of instrumental rationality. Ariel Rubinstein puts this point well when discussing the value of formal economic models:

> Formal language imposes self-discipline on the storyteller. A teller of economic tales who uses formal language is obliged to spell out his assumptions precisely. [...] A description of an economic model is like the introduction in a tale, presenting the heroes, their interests and the setting in which they operate. An array of rules by which the model is "allowed" to develop from its beginning to its end is called a *solution concept*. (2012, 19, original emphasis)

Moehler asks us to accept that the WPU is a uniquely rational principle of conflict resolution on the basis of a tale about *homo prudens* trying to resolve conflicts of interest in a "peace game". We use game theory to clarify key elements of that tale; we use the solution concept of subgame perfect equilibrium to ensure that the tale unfolds in a manner consistent with the standard theory of rationality. The modeling exercise thus helps us expose the ways in which the plausibility of Moehler's complex contractarian tale relies on subtle but significant departures from the conventional dictates of rational prudence.

REFERENCES

Binmore, Ken. 1998. *Game Theory and the Social Contract: Just Playing*. Cambridge, MA: MIT Press.

Buchanan, James M. 2000. *The Limits of Liberty: Between Anarchy and Leviathan*. The Collected Works of James M. Buchanan Indianapolis: Liberty Fund.

Fudenberg, Drew and Jean Tirole. 1991. *Game Theory*. Cambridge, MA and London: MIT Press.

Gaus, Gerald. 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. New York: Cambridge University Press.

Gaus, Gerald. 2019. "Moral Conflict and Prudential Agreement: Michael Moehler's *Minimal Morality*." *Analysis* 79(1):106–115.

Gauthier, David P. 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Hobbes, Thomas. 1985. *Leviathan*. Indianapolis: Hackett.

Moehler, Michael. 2010. "The (Stabilized) Nash Bargaining Solution as a Principle of Distributive Justice." *Utilitas* 22(4):447–473.

Moehler, Michael. 2014. "The Scope of Instrumental Morality." *Philosophical Studies* 167(2):431–451.

Moehler, Michael. 2018. *Minimal Morality: A Multilevel Social Contract Theory*. Oxford: Oxford University Press.

Moehler, Michael. forthcoming. "Diversity, Stability, and Social Contract Theory." *Philosophical Studies* .

Rodrik, Dani. 2015. *Economics Rules: Why Economics Works, When It Fails, and How To Tell The Difference*. Oxford: Oxford University Press.

Rubinstein, Ariel. 1982. "Perfect Equilibrium in a Bargaining Model." *Econometrica* 50(1):97–109.

Rubinstein, Ariel. 2012. *Economic Fables*. Cambridge: Open Book Publishers.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Sugden, Robert. 1990. "Contractarianism and Norms." *Ethics* 100(4):768–86.

Thoma, Johanna. 2015. "Bargaining and the Impartiality of the Social Contract." *Philosophical Studies* 172(12):3335–3355.

Thrasher, John. 2013. "Uniqueness and Symmetry in Bargaining Theories of Justice." *Philosophical Studies* 167(3):683–99.

6. APPENDIX: PROOFS

This appendix proves statements in the text about various equilibria for the original peace game (i.e., a repeated version of the stage game in fig. 1). Recall that our analysis assumes $V > 0$, $V > d_i$, and $W_i(V) > d_i$ for $i \in \{R, C\}$.

Proposition 1. *Given our assumptions, coordinating on the WPU is a subgame perfect equilibrium.*

Proof. Assume Collins plays Unconditional WPU and consider an arbitrary stage t from Rowan's perspective. If Rowan deviates to ROW or COL at t and then returns to playing Unconditional WPU, his expected payoff is $d_R + \frac{\delta W_R(V)}{1-\delta}$; if he stays with WPU at t and plays Unconditional WPU thereafter, his expected payoff is $W_R(V) + \frac{\delta W_R(V)}{1-\delta}$. Since $W_R(V) > d_R$, Rowan maximizes expected utility by playing Unconditional WPU and, thus, has no incentive to deviate from Unconditional WPU at any stage. By similar reasoning (given the symmetry of the game), Collins plays Unconditional WPU if Rowan plays Unconditional WPU. □

Proposition 2. *If, in addition to our initial assumptions, $d_i > 0$ for $i \in \{R, C\}$, then there is a conflict equilibrium in which both players demand everything all the time.*

Proof. Assume Collins plays Unconditional COL. Then, at any arbitrary stage t, Rowan's payoff for playing COL is $0 + \frac{\delta d_R}{1-\delta}$, while his payoff for playing ROW or WPU is $d_R + \frac{\delta d_R}{1-\delta}$. Since we're assuming $d_R > 0$, Rowan does best by playing ROW or WPU (and is indifferent between the two) at every stage t and, thus, by playing Unconditional ROW or Unconditional WPU. So Unconditional ROW is a best response to Unconditional COL. By analogous reasoning, we can show that Collins has no incentive to deviate from Unconditional COL assuming Rowan plays Unconditional ROW. □

Proposition 3. *If, in addition to our initial assumptions, $d_i < 0$ for $i \in \{R, C\}$, then (i) coordinating on COL is a subgame perfect equilibrium, and (ii) coordinating on ROW is a subgame perfect equilibrium.*

Proof. Assume Collins plays Unconditional COL. Then, at any arbitrary stage t, Rowan's payoff for play COL is 0, while his payoff for playing ROW or WPU is $d_R + 0$. Since we're assuming $d_R^t < 0$, Rowan does best by playing COL at every stage t and, thus, by playing Unconditional COL. By similar reasoning, Collins plays Unconditional COL if Rowan plays Unconditional COL. (An analogous argument shows that coordinating on ROW is an equilibrium.) □